Small-molecule Bioactivity Databases

CHAPTER 16

Small-molecule Bioactivity Databases

SEAN EKINS,*[a,b] ALEX M. CLARK,[a,c] CHRISTOPHER SOUTHAN,[d] BARRY A. BUNIN[a] AND ANTONY J. WILLIAMS[e]

[a] Collaborative Drug Discovery, Inc., 1633 Bayshore Highway, Suite 342, Burlingame, CA 94010, USA; [b] Collaborations Pharmaceuticals, Inc., 5616 Hilltop Needmore Road, Fuquay-Varina, NC 27526, USA; [c] Molecular Materials Informatics, Inc., 1900 St. Jacques #302, Montreal H3J 2S1, Quebec, Canada; [d] IUPHAR/BPS Guide to PHARMACOLOGY, Centre for Integrative Physiology, University of Edinburgh, Hugh Robson Building, Edinburgh, EH8 9XD, UK; [e] ChemConnector, 513 Chestnut Grove Court, Wake Forest, NC 27587, USA
*Email: ekinssean@yahoo.com

16.1 Introduction

Over the last decade there has been a proliferation of chemistry databases on the internet.[1,2] We have gone from a point in the early 2000's when there was little in the way of small-molecule and bioactivity data available online, to today, where web based publicly accessible databases can contain tens of millions of molecules. Many of these databases have over a million bioactivity data points [such as half-maximal inhibitory concentration (IC_{50}) or inhibitor binding affinity (K_i)[2]] and data are shared and proliferated between them (*e.g.* ChEMBL, https://www.ebi.ac.uk/chembl/, PubChem, https://pubchem.ncbi.nlm.nih.gov/, and other databases mirror some of each other's data). The evolution of these bioactivity databases has followed different routes. Examples include collections of molecules with one or more

Chemical Biology No. 1
High Throughput Screening Methods: Evolution and Refinement
Edited by Joshua A. Bittker and Nathan T. Ross

Published by the Royal Society of Chemistry, www.rsc.org

particular related bioactivity, collections of multiple curated sets of data, user deposited datasets and combinations of these. Databases were once mainly used to look up structure and properties, and as they expanded to include experimental and predicted properties their function shifted. Increasingly, these databases are used to predict potential targets based on the structure similarity principle,[3–5] chemical–biological read-across[6] and toxicology profiling,[7,8] and in many ways have evolved into portals for different data type.

In parallel, commercial databases, such as Chemical Abstracts (CAS) SciFinder®[9] and GVKBio, focused on curated chemical structures, some of which have been quantitatively assessed for their complementarity with public databases and found to contain unique content.[10–12] We have previously discussed the potential for divergence of these commercial systems from the public databases.[2] The focus of this chapter will be on freely accessible databases such as BindingDB (www.bindingdb.org), PubChem, ChEMBL, International Union of Basic and Clinical Pharmacology (IUPHAR)/BPS Guide to PHARMACOLOGY (GtoPdb, http://guidetopharmacology.org/) and public data in the Collaborative Drug Discovery (CDD) Vault. We also refer readers to earlier publications and discussions regarding public domain compound databases that have covered other systems and content.[13–17]

There have been numerous comparisons of public bioactivity databases at the level of molecules or targets that have suggested complementarity, and we do not intend to add any more from this perspective.[18] There have also been efforts to combine different bioactivity databases. For example, Confederated Annotated Research Libraries of Small Molecule Biological Activity Data (CARLSBAD) brought together ChEMBL, GtoPdb, PubChem, WOMBAT[19] and PDSP (http://kidbdev.med.unc.edu/databases/kidb.php)[20] in order to help facilitate chemical biology research and data mining.[21] CARLSBAD (http://carlsbad.health.unm.edu) is only available to academics and non-commercial researchers; and even then one must apply in order to access it, which would likely deter the casual user. Another example of such a combined database is the ChemProt database,[22] which is made up of data from seven databases and contains 1.7 million compounds and 7.8 million bioactivity measurements. It uses Daylight like fingerprints and can calculate the similarity ensemble approach (SEA).[23] A naive Bayesian classifier was used with the Daylight like and Morgan fingerprints to build models for 850 proteins. Performance was described for only one model for hERG, although models for 143 other proteins were also suggested to outperform SEA.[22]

16.2 Public Bioactivity Databases

There are now likely tens and possibly hundreds of bioactivity databases available online or for download, many of which are unknown to the general audience and perhaps only accessible as supplemental data in publications. If you can imagine that a collection of molecules can be curated and used for a single paper, then classed as a database and made available as

supplemental data, then that would give some idea of the scope of bioactivity databases. For example, approximately 1000 natural products from African medicinal plants have been collated, analyzed and made available.[24] More extensive datasets with a small number of molecules but hundreds of assays represent a rich data source. One example is the ToxCast™ project, which was launched in 2007 and is a long term, multi-million dollar effort that hopes to understand biological processes impacted by chemicals that may lead to adverse health effects, as well as generate predictive models that should enable predictions of toxicity.[25] The project is a multi-phase project and, currently in phase 3, it covers over 3800 unique chemicals and up to 900 assays, including nuclear receptors, *etc.*[26] The phase 1 and 2 data have been made available *via* the ToxCast dashboard (http://actor.epa.gov/dashboard/) and are available in various forms for download,[27] and therefore, can be meshed into other databases for the toxicology community. Phase 3 is presently underway and the data will be released in phases throughout the lifetime of this part of the project. ToxCast data are not presently available *via* PubChem. A related long term project is Tox21 (https://www.epa.gov/chemical-research/toxicology-testing-21st-century-tox21),[28] a collaboration between the US National Institutes of Health (NIH), Food and Drug Administration (FDA) and Environmental Protection Agency (EPA). Tox21 data are generally made available *via* PubChem,[29] albeit with a staged release cycle.

A major challenge in this area is how to curate all of these individual bioactivity datasets and databases, which may only be accessible *via* formats that are custom designed for the task (*e.g.* SQL database dumps or CSV files) or have formats such as MDL SDfile that lack provenance. Aggregating individual datasets let alone databases is a complex task with potential issues in standardization or normalization of data across sources, duplication of data and structure, as well as identification of errors, *etc*. While there have been some efforts directed towards automation of data curation, heavy emphasis on manual curation is likely to be needed to resolve conflicts. Databases such as ZINC assemble and host the chemistry related features of many of these SDF files in a single place, but are not known as bioactivity databases. Small focused databases, such as chemical modulators of epigenome reader domains (ChEpiMod; http://chepimod.org/), which combine data from manual extraction of publications and patents, as well as databases such as ChEMBL,[30] focus on domains rather than proteins. Another example is GLASS, which focuses on G-protein coupled receptor (GPCR) ligands collected from ChEMBL, BindingDB, GtoPdb, DrugBank[31] and PDSP,[20] and currently has over 276 000 unique ligands and over 3000 GPCRs (http://zhanglab.ccmb.med.umich.edu/GLASS/).[32] These are just a small sample of additional bioactivity datasets and databases in a variety of formats. The utility of many bioactivity databases includes simple look ups for information relating to a structure series of interest through to building structure activity models. We will now give a brief summary of several well-known public bioactivity databases and how they might be used. While there are certainly other examples of early databases, such as ChemBank launched

in 2003,[33–35] among others,[36,37] the following are notable for their continued influence and size.

16.2.1 BindingDB

The BindingDB started following a 1997 workshop on the need for a database focused on binding thermodynamics that could capture binding affinities, experimental details, facilitate a wide range of queries, be publically accessible and allow user deposition.[38] The database was launched in 2000. At the time of writing, BindingDB hosts over 1 207 821 binding data, 6265 protein targets and 529 618 molecules. These can be used in various ways, including considering off-target activity, target prediction, finding compounds for targets (Figure 16.1), virtual screening and structure activity modeling, *etc*.[39] BindingDB uses the SEA approach[23,40–42] to rank targets. While the database collects data from other sources, the deposition and manual curation allows for error checking and correction. Reuse of the data is less restrictive than with ChEMBL (see below). BindingDB processes extensive amounts of data from ChEMBL but organizes it in a way that offers users complementary options for interrogating the content of both resources.

16.2.2 PubChem

At the time of writing, PubChem[43] contained 89 124 111 compounds and 1 154 429 bioassays with 229 972 149 bioactivities for 2 101 164 tested compounds,[44,45] making it the largest free online bioactivity database. It was initially launched by the NIH in 2004 to support the ''New Pathways to Discovery'' component of their roadmap initiative.[44] The primary purpose for the database was to act as a *repository* for the Molecular Libraries Screening Centers Network (MLSCN) screening results that were expected to yield chemical probes. Clearly, it now extends well beyond this and encompasses all of the screening data behind these probe hunting efforts as well as hosting data mirrored from other databases such as ChEMBL. In some ways, focusing on additional information has neglected the probes, making those specific data hard to find.[46] A new derivative database of PubChem is called the BioAssay Research Database (BARD; https://bard.nih.gov), and is used for housing screening data for probe development. BARD was released in 2015 and uses a controlled vocabulary to describe the assay protocols, enabling more structured and automated bioactivity analysis. A limitation of this free database is that the backend relies on several software components that require licenses,[47] limiting local deployment possibilities. PubChem has built itself up to become the definitive bioactivity database in terms of scale, public accessibility, and the ease of a quick look up for a compound and potential bioactivity (Figure 16.2). Similar to other large submission based resources, it has been criticized for allowing the submission of vendor libraries, including ''make on demand'' compounds

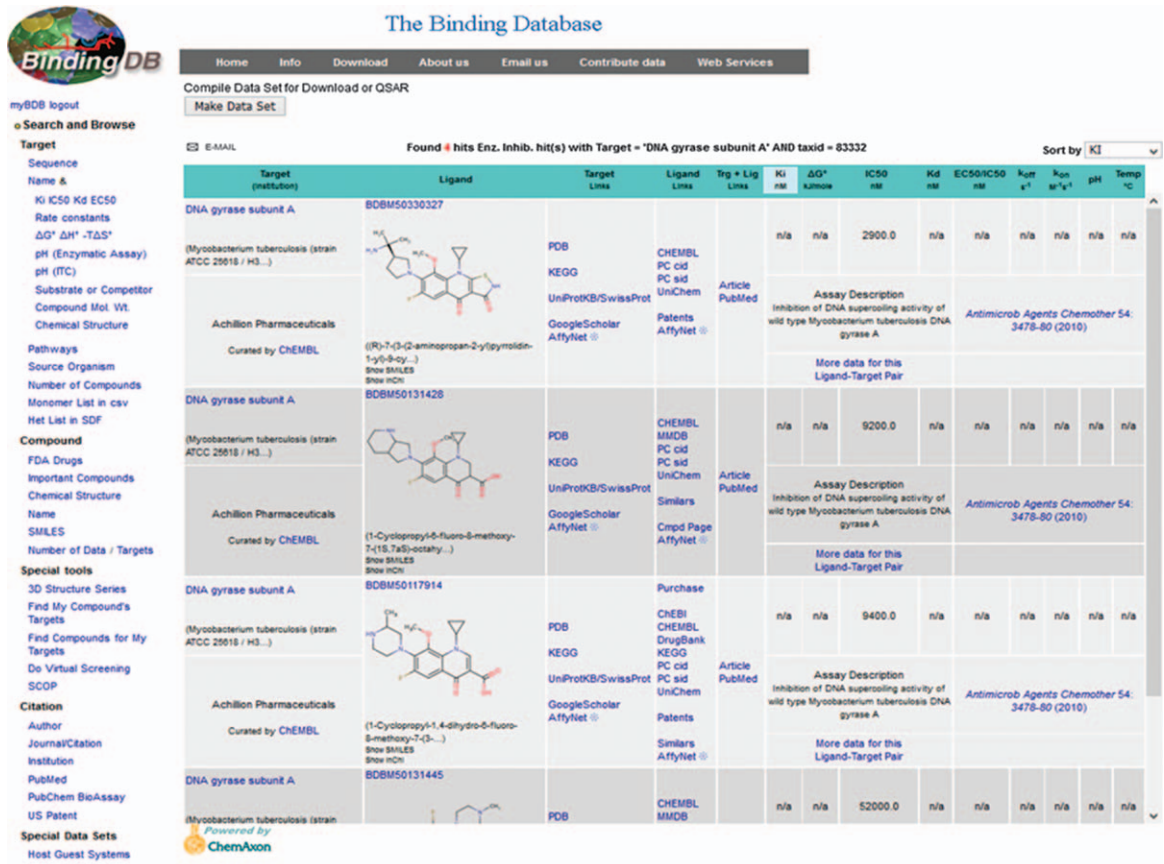


Figure 16.1 An overview of the BindingDB database showing some DNA gyrase inhibitors for *Mycobacterium tuberculosis.* DOI: 10.6084/m9.figshare.3206236.

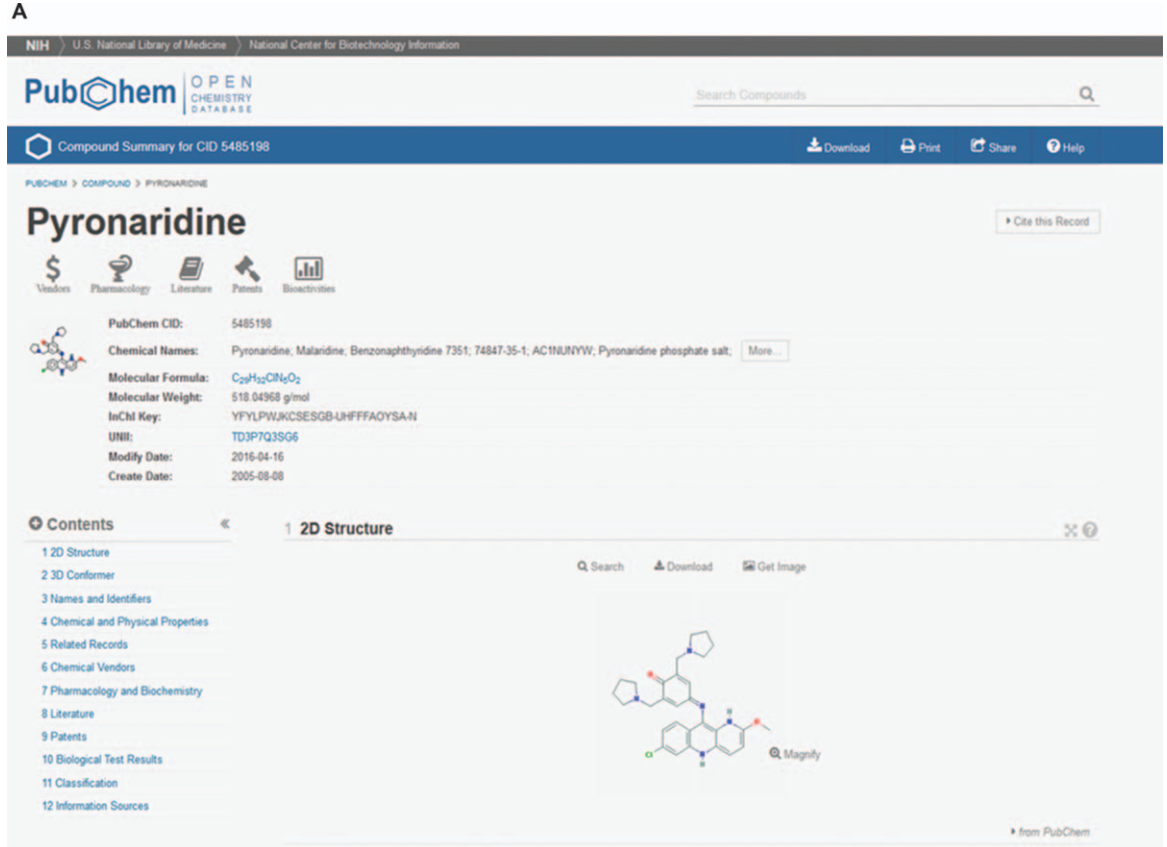


Figure 16.2 (A) An example of a compound summary page on PubChem showing pyronaridine. (B) Pubchem bioactivity data for pyronaridine. DOI: 10.6084/m9.figshare.3206236.

Figure 16.2 Continued.

that had never actually existed and certainly had never been screened in bioassays (indeed the largest of these vendor submitters has now been restricted to stock-only records). PubChem has now become a host for general chemical data, serving many needs, including access to safety data[48] and a variety of other compound data that can be viewed inside the PubChem Classification Browser.[49] PubChem has made specific efforts to ensure accessibility by the provision of so-called ''widgets'',[50] which have been used in the recently released EPA iCSS Chemistry Dashboard (http://comptox.epa.gov/) to provide direct access from their platform to an embedded view of the PubChem data. While departing from its original mandate, this added scope positions PubChem, along with others such as ChemSpider,[51] as a provider of a valuable community service across chemistry applications. Databases such as PubChem have considerable utility for predicting compound–target associations, including one example describing a bioactivity profile similarity search (BASS).[52] While not necessarily novel, the sheer volume of data now accessible puts these types of approaches within reach of scientists in academia or small companies. The available data can be utilized by users to build their own quantitative structure–activity relationship or machine learning models, or can be searched in order to propose similar compounds that can then be tested in other assays. This fundamental shift is based on the available data, most of which it is hoped will be released under open data licenses.[53] However, more investment is needed for utilization training and awareness.

16.2.3 ChEMBL

ChEMBL is a database of drugs and other small molecules of biological interest.[54–56] ChEMBL_21 contains 1 592 191 compounds with 13 967 816 activities. It includes target binding relationships for small molecules, the effect of these compounds on cells and organisms (*e.g.* K_i, IC_{50}), and associated absorption, distribution, metabolism and excretion (ADME)/toxicity (Tox) data. In contrast to PubChem, ChEMBL has focused specifically on literature extraction, but since 2011 it has also included a filtered subset of confirmatory PubChem BioAssay results.[57] The database contains manually curated structure–activity relationship (SAR) data from the primary medicinal chemistry and pharmacology literature, and therefore, provides high quality data that may be used for computational purposes. As described herein, it has been used extensively for SEA analyses as well as data aggregation efforts. In addition, ChEMBL data have been used to assess the reproducibility of kinase selectivity studies.[58] Data for the rat and human adenosine receptors from ChEMBL have been used to perform virtual screening based on proteochemometric modeling, resulting in the identification of novel inhibitors.[59] ChEMBL has obtained large datasets from industry for neglected diseases such as malaria[60] and ADME/Tox datasets that AstraZeneca have published[61] (Figure 16.3). The focused portals it has created include ChEMBL-Neglected Tropical Disease (NTD), Kinase SARfari, GPCR SARfari[62] and ADME SARfari.


A
ChEMBL
EBI > Databases > Small Molecules > ChEMBL Database
ChEMBL
Downloads
UniChem
Malaria Data
ChEMBL-NTD
ADME SARfari New
Kinase SARfari
GPCR SARfari
DrugEBIlity
Document Report Card
Doc ID CHEMBL3301361
Title Experimental in vitro DMPK and physicochemical data on a set of publicly disclosed compounds
Authors Mark Wenlock and Nicholas Tomkinson
Abstract Experimental data determined at AstraZeneca on a set of compounds in the following assays: pKa, lipophilicity (LogD7.4), aqueous solubility, plasma protein binding (human, rat, dog , mouse and guinea pig), intrinsic clearance (human liver microsomes, human and rat hepatocytes). The references provided for the assays exemplify the experimental procedures used in generating the data.
DOI http://dx.doi.org/10.6019/CHEMBL3301361

B

ChEMBL Assay ID	Assay Source	Assay Type	Assay Organism	Description	Activity Count	Reference	
CHEMBL3301369	AstraZeneca Deposited Data	A		% bound to plasma by equilibrium dialysis. Compound is incubated with whole guinea pig plasma at 37C for >5hrs. Method described in B. Testa et al (Eds.), Pharmacokinetic Profiling in Drug Research: Biological, Physicochemical, and Computational Strategies, Wiley-VCH, Weinheim, 2006, pp.119-141. Experimental range 10% to 99.95% bound.	91	CHEMBL3301361	☑
CHEMBL3301367	AstraZeneca Deposited Data	A		% bound to plasma by equilibrium dialysis. Compound is incubated with whole dog plasma at 37C for >5hrs. Method described in B. Testa et al (Eds.), Pharmacokinetic Profiling in Drug Research: Biological, Physicochemical, and Computational Strategies, Wiley-VCH, Weinheim, 2006, pp.119-141. Experimental range 10% to 99.95% bound.	244	CHEMBL3301361	☑
CHEMBL3301366	AstraZeneca Deposited Data	A		% bound to plasma by equilibrium dialysis. Compound is incubated with whole rat plasma at 37C for >5hrs. Method described in B. Testa et al (Eds.), Pharmacokinetic Profiling in Drug Research: Biological, Physicochemical, and Computational Strategies, Wiley-VCH, Weinheim, 2006, pp.119-141. Experimental range 10% to 99.95% bound.	717	CHEMBL3301361	☑
CHEMBL3301365	AstraZeneca Deposited Data	A		% bound to plasma by equilibrium dialysis. Compound is incubated with whole human plasma at 37C for >5hrs. Method described in B. Testa et al (Eds.), Pharmacokinetic Profiling in Drug Research: Biological, Physicochemical, and Computational Strategies, Wiley-VCH, Weinheim, 2006, pp.119-141. Experimental range 10% to 99.95% bound.	1614	CHEMBL3301361	☑
CHEMBL3301368	AstraZeneca Deposited Data	A		% bound to plasma by equilibrium dialysis. Compound is incubated with whole mouse plasma at 37C for >5hrs. Method described in B. Testa et al (Eds.), Pharmacokinetic Profiling in Drug Research: Biological, Physicochemical, and Computational Strategies, Wiley-VCH, Weinheim, 2006, pp.119-141. Experimental range 10% to 99.95% bound.	162	CHEMBL3301361	☑
CHEMBL3301370	AstraZeneca Deposited Data	A		Intrinsic clearance measured in human liver microsomes following incubation at 37C. Experimental range <3 to >150 microL/min/mg. Rapid Commun. Mass Spectrom. 2010, 24, 1730-1736.	1102	CHEMBL3301361	☑
CHEMBL3301371	AstraZeneca Deposited Data	A		Intrinsic clearance measured in rat hepatocytes following incubation at 37C. Experimental range <3 to >150 microL/min/1E6 cells. Rapid Commun. Mass Spectrom. 2010, 24, 1730-1736	837	CHEMBL3301361	☑
CHEMBL3301372	AstraZeneca Deposited Data	A		Intrinsic clearance measured in human hepatocytes following incubation at 37C. Experimental range <3 to >150 microL/min/1E6 cells. Rapid Commun. Mass Spectrom. 2010, 24, 1730-1736.	408	CHEMBL3301361	☑

Figure 16.3 Examples of the data in ChEMBL. (A) AstraZeneca *in vitro* data report card. (B) Detail on the individual AstraZeneca *in vitro* DMPK datasets.
Credit: European Bioinformatics Institute. DOI: 10.6084/m9.figshare.3206236.

16.2.4 GtoPdb

GtoPdb is the successor of an earlier database, IUPHAR-DB, which was focused on receptors and channels mapped to endogenous ligands, clinical drug candidates and research compounds. This was established in 2009 under the auspices of the IUPHAR Committee on Receptor Nomenclature and Drug Classification (NC-IUPHAR).[42,43] From 2012 to 2015 Wellcome Trust funding facilitated the expansion of IUPHAR-DB into GtoPdb, which now covers all human pharmacological target classes. Release 2016.2 covers 2775 targets, 8400 ligands, 14 327 binding constants and 29 247 references. The website features extensive curated information and links. As an example, searching for the histone deacetylase 2 (HDAC2) target retrieves gene and protein information, database links (Figure 16.4A), as well as a downloadable list of inhibitors (Figure 16.4B). Note that approved drugs, ligand selectivity and small-molecule status are highlighted (Figure 16.4C). This resource has two other unique features. The first is the support of the NC-IUPHAR target class sub-committees for content selection. The second is collaboration with the British Journal of Pharmacology for the publication of biannual content overviews, live tables of links and instructions for authors to get their results ready for curation.[63,64]

16.2.5 Public Data in the CDD Vault

In 2004, CDD (https://www.collaborativedrug.com/) started to develop the CDD Vault[65] as a web based database that would enable scientists to move away from storing their data in spreadsheets, and make them accessible to mining and sharing in their group or with collaborators from any browser. A number of applications for collaborative research have previously been described[66] regarding the large number of public datasets in the CDD Vault and their use for dataset analysis.[67] Public data in the CDD Vault can be accessed by anyone after first registering (http://www.collaborativedrug.com/register) and data can be searched across over 100 datasets. Currently, these cover several vendor libraries as well as unique datasets submitted by researchers and companies. CDD has a considerable focus on datasets for screening against neglected diseases such tuberculosis,[17,68–81] malaria,[82] Chagas disease[83] and Ebola.[84] Many of these datasets have been used in the drug discovery efforts of those submitting them. In addition, CDD has included other datasets and then enhanced them. For example, the physicochemical property datasets deposited by AstraZeneca in ChEMBL (Figure 16.5A) have been used in the CDD Vault and the data visualized alongside calculated properties (Figure 16.5B). These efforts perhaps point to some of the dataset and database integration challenges. For example, CDD Public contains a small fraction of the datasets from ChEMBL or PubChem, and has focused on very specific areas such as neglected diseases. Creating long lists of datasets is not ideal and the organization of them by type is currently rudimentary. Efforts to use the CDD Vault to host the

currently available bioactivity databases described above have not been attempted but could be valuable.

While the public datasets in the CDD Vault showcase how many different datasets can be searched, the CDD Vault can also archive the user's own data and allow them to mine a broad range of diverse objects that can later be selectively and securely shared with other researchers (or permanently kept private, which is the default behavior). The CDD web based database architecture handles a broad array of data types (*e.g.* CSV and SD file convertible formats that represent the chemical and biological data) and incorporates industry standard Marvin chemical structure tools, calculator plug-ins for physicochemical calculations and the JChem Cartridge for structure searching from ChemAxon (Budapest, Hungary). These features

A

Gene and Protein Information

Species	TM	AA	Chromosomal Location	Gene Symbol	Gene Name	Reference
Human	-	488	6q21	*HDAC2*	histone deacetylase 2	
Mouse	-	488	10 B1	*Hdac2*	histone deacetylase 2	
Rat	-	-	20q12	*Hdac2*	histone deacetylase 2	

Previous and Unofficial Names

RPD3

YAF1

Database Links

BRENDA	3.5.1.98
ChEMBL Target	CHEMBL1937 (Hs)
DrugBank Target	Q92769 (Hs)
Ensembl Gene	ENSG00000196591 (Hs), ENSMUSG00000019777 (Mm), ENSRNOG00000000604 (Rn)
Entrez Gene	3066 (Hs), 15182 (Mm), 84577 (Rn)
ExplorEnz	3.5.1.98
GeneCards	*HDAC2* (Hs)
GenitoUrinary Development Molecular Anatomy Project	*Hdac2* (Mm)
InterPro	Q92769 (Hs), P70288 (Mm)
KEGG Enzyme	3.5.1.98
KEGG Gene	hsa:3066 (Hs), mmu:15182 (Mm), rno:84577 (Rn)
NeXtProt	Q92769 (Hs)
OMIM	605164 (Hs)
RefSeq Nucleotide	NM_001527 (Hs), NM_008229 (Mm), NM_053447 (Rn)
RefSeq Protein	NP_001518 (Hs), NP_032255 (Mm), NP_445899 (Rn)
TreeFam	*HDAC2*
UniGene Hs.	3352 (Hs)
UniProtKB	Q92769 (Hs), P70288 (Mm)
Wikipedia	*HDAC2* (Hs)

Enzyme Reaction

EC Number: 3.5.1.98

Download all structure-activity data for this target as a CSV file

Figure 16.4 (A) Summary of a search for HDAC2 in Guide to Pharmacology. (B) List of HDAC2 inhibitors. (C) Compound selectivity profile against other HDACs. DOI: 10.6084/m9.figshare.3206236.

B

Inhibitors

Key to terms and symbols | View all chemical structures | Click column headers to sort

Ligand	Sp.	Action	Affinity	Units	Reference
romidepsin	Hs	Inhibition	10.4	pK_i	2
apicidin	Hs	Inhibition	9.9	pK_i	2
trichostatin A	Hs	Inhibition	9.2	pK_i	2
belinostat	Hs	Inhibition	9.1	pK_i	2
dacinostat	Hs	Inhibition	8.9	pK_i	2
vorinostat	Hs	Inhibition	8.8	pK_i	2
scriptaid	Hs	Inhibition	8.7	pK_i	2
givinostat	Hs	Inhibition	8.5	pK_i	2
mocetinostat	Hs	Inhibition	7.5	pK_i	2
entinostat	Hs	Inhibition	7.2	pK_i	2
tacedinaline	Hs	Inhibition	6.8	pK_i	2
panobinostat	Hs	Inhibition	8.5	pEC_{50}	5
givinostat	Hs	Inhibition	7.3	pEC_{50}	5
apicidin	Hs	Inhibition	6.9	pEC_{50}	5
belinostat	Hs	Inhibition	6.9	pEC_{50}	5
entinostat	Hs	Inhibition	5.9	pEC_{50}	5
santacruzamate A	Hs	Inhibition	9.9	pIC_{50}	7
quisinostat	Hs	Inhibition	9.5	pIC_{50}	1
CHR-3996	Hs	Inhibition	8.4	pIC_{50}	6
CUDC-907	Hs	Inhibition	8.3	pIC_{50}	8
CUDC-101	Hs	Inhibition	7.9	pIC_{50}	3
ricolinostat	Hs	Inhibition	7.3	pIC_{50}	9
butyric acid	Hs	Inhibition	4.9	pIC_{50}	4

Inhibitor Comments

Vorinostat has high affinity for HDACs 2, 3, 6, 9, 10 and 11, but 10-fold lower affinity for HDAC8.

General Comments

HDAC2 is a Class I histone deacetylase.

Figure 16.4 Continued.

allow similarity and substructure searching, and more complex analyses within the application. The CDD Vault is used as the database behind several large collaborative projects such as NIH Blueprint, Bill and Melinda Gates Foundation Tuberculosis Drug Accelerator, Kinetoplastid drug development consortium and More Medicines for Tuberculosis, with each sharing data in different ways in their own secure environment. CDD has therefore enabled complex collaborations to become manageable and scalable using their technologies.

16.2.5.1 CDD Models in the CDD Vault

The capacity to build Bayesian models with open source ECFP6 and FCFP6 fingerprints (https://github.com/cdd/modified-bayes)[85] is available in the CDD Vault and implemented as CDD Models. This provides a powerful machine learning technology to scientists that can be used in a secure CDD Vault to build and share models.[86] This work built on earlier efforts with collaborators at Pfizer to show that open source tools could produce

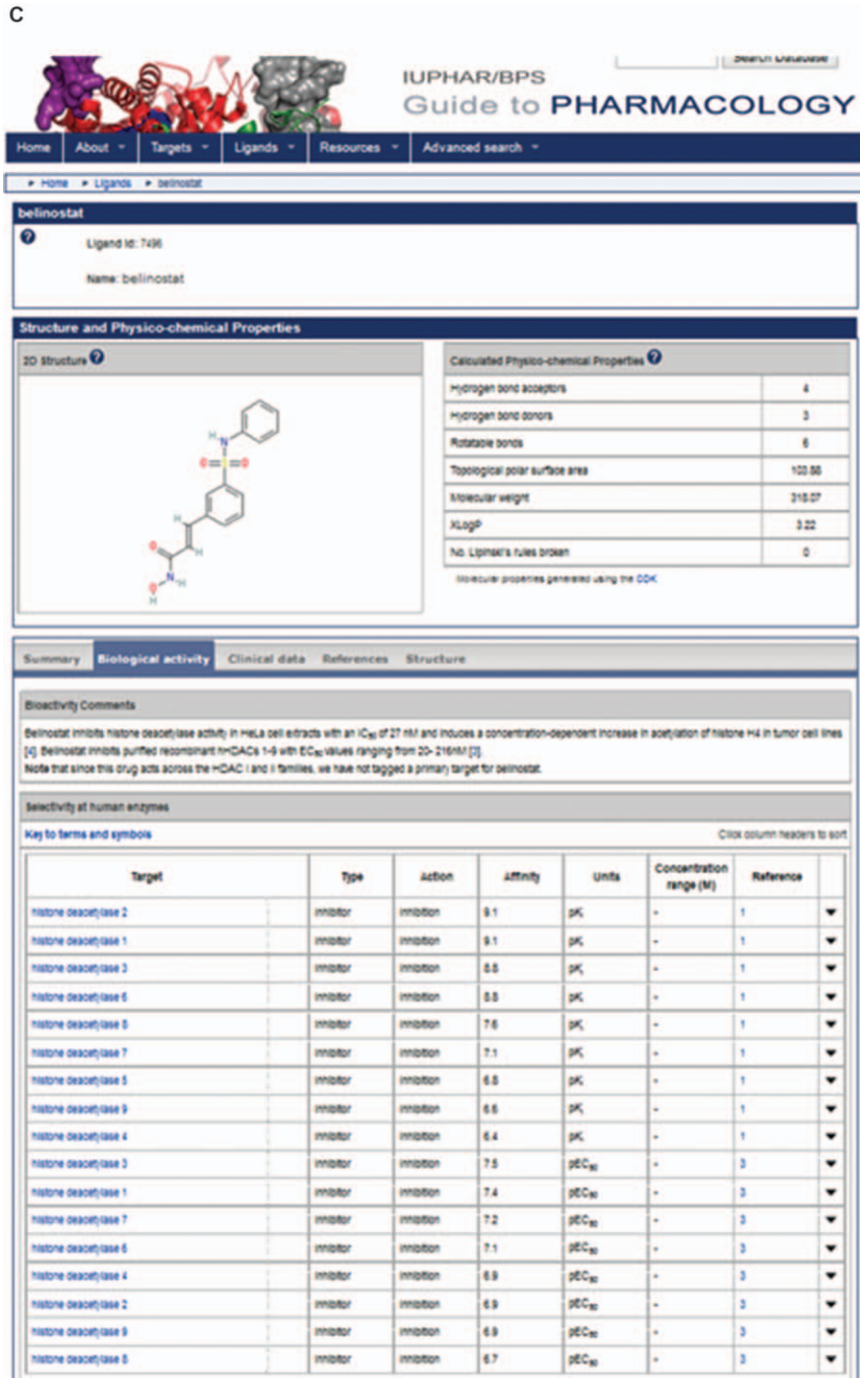
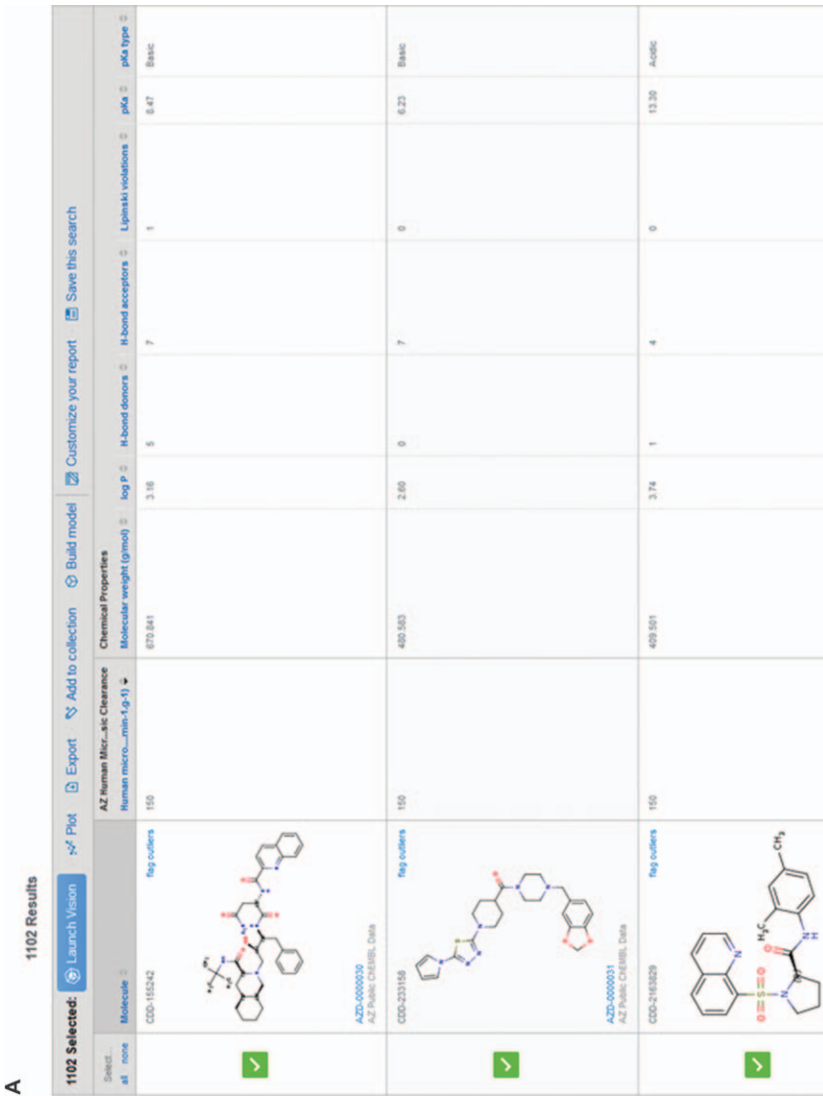


Figure 16.4 Continued.

A


1102 Results
1102 Selected:
Launch Vision
Plot
Export
Add to collection
Build model
Customize your report
Save this search
Select: all none
Molecule
AZ Human Micr...sic Clearance
Human micro....min-1.g-1)
Chemical Properties
Molecular weight (g/mol)
log P
H-bond donors
H-bond acceptors
Lipinski violations
pKa
pKa type
CDD-155242
flag outliers
150
670.841
3.16
5
7
1
8.47
Basic
AZD-0000030
AZ Public ChEMBL Data
CDD-233158
flag outliers
150
480.583
2.60
0
7
0
6.23
Basic
AZD-0000031
AZ Public ChEMBL Data
CDD-2163829
flag outliers
150
409.501
3.74
1
4
0
13.30
Acidic

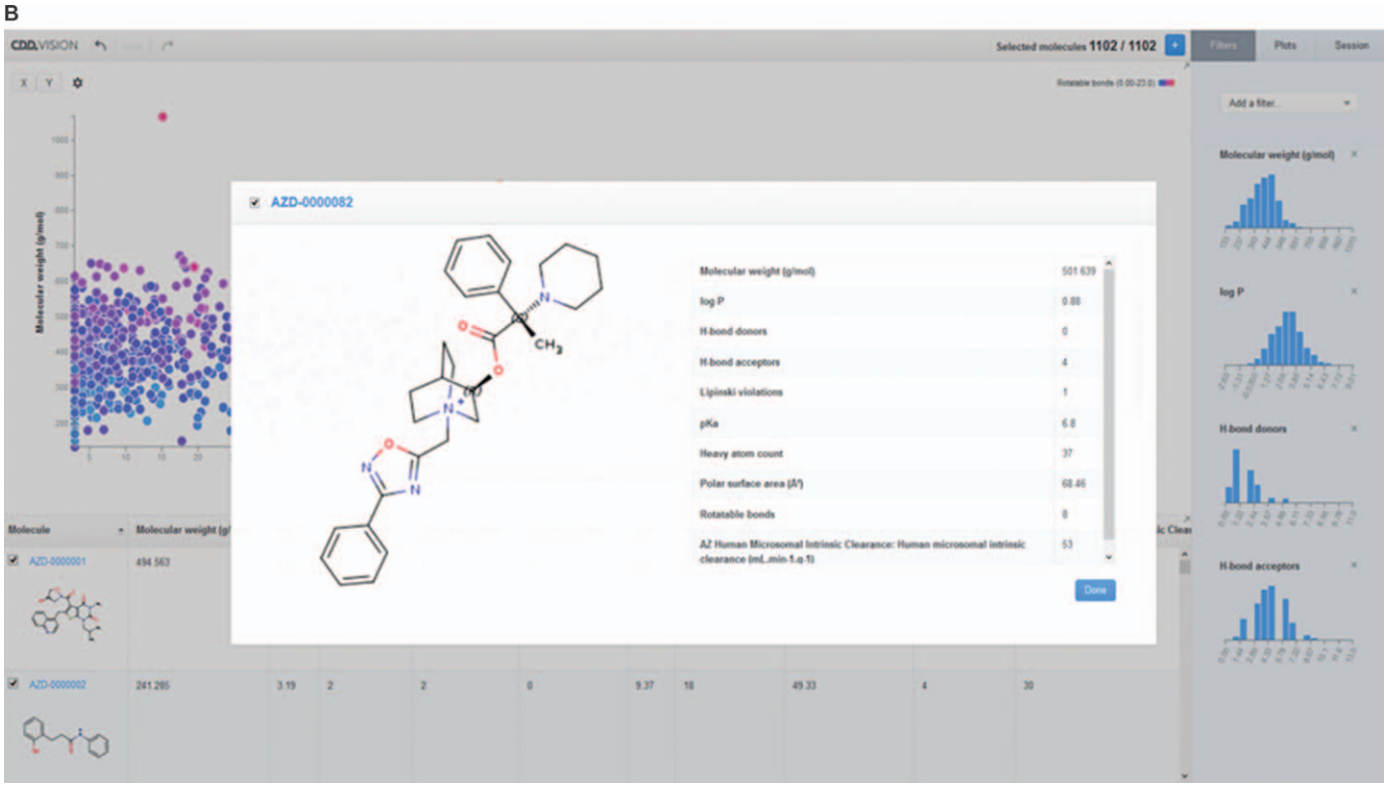


Figure 16.5 An example of a dataset in the CDD Vault. (A) A spreadsheet view and (B) CDD Vision view of AstraZeneca microsomal intrinsic clearance. DOI: 10.6084/m9.figshare.3206236.

comparable models to those generated with commercial tools when modeling ADME data.[87] It has been shown that commercial fingerprint descriptors and Bayesian models could be used to pick active compounds after virtual screening, with good enrichments and hit rates for *Mycobacterium tuberculosis*,[73,75,88–90] predict *in vivo Mtb* activity in mice,[80] and be used to identify leads and repurpose drugs for Chagas disease[83] and Ebola.[84] We have applied this machine learning approach to modeling decision making for chemical probes,[8] ADME-Tox models[9] as well as microsomal stability in mice.[91] As an example, we have used the public AstraZeneca physicochemical property and ADME data to build models (Figure 16.6). The open source descriptors and Bayesian algorithm have also been used outside of the CDD Vault to create several thousand Bayesian models with the ChEMBL data[10] or manually curated data from other sources.[92] One example of the utility of such ChEMBL data involved cleaning up and using the data to create a Bayesian model of 536 HDAC2 inhibitors to produce models with excellent receiver operating characteristic (ROC) values (>0.89; Figure 16.7). The Bayesian approach is undergoing continual refinement, most recently with a Bayesian binning approach.[93] By enabling model building in the CDD Vault we have gone someway to creating a machine learning model repository. While there are academic efforts in this area,[94,95] CDD Models may represent the first commercial effort, and this aspect could be expanded further, creating a database that allows the user to flip between models and the data underpinning them. Until then, we have created thousands of models and made them accessible through web pages (Table 16.1).

In order to test some of the open technologies created we have opted to prototype them in a mobile app called TB Mobile.[85,96] This app can be thought of as a subset of one of the public datasets in the CDD Vault relating to compounds and targets.[72] We first demonstrated the use of the fingerprints and Bayesian algorithms in this app to predict potential targets for compounds in addition to using similarity calculations and clustering of data. Such apps themselves could be used and considered as bioactivity databases, although it remains to be seen whether more will be created like them, and what challenges and benefits these in turn will create as they may represent silos that cannot be readily integrated.

16.3 Data Quality

Within the global cheminformatics community concerns have surfaced in recent years over the quality of data in public chemistry and bioactivity databases. Initially our focus was on data released by drug companies and how the quality of the compounds compared from the perspective of physicochemical properties and reactive groups.[97] We then turned our attention to ''new databases'' as they were released and found frequent issues in the curation of molecule structures,[13] which in turn led us to larger scale analysis of many public databases and the analysis of the proliferation of errors

Figure 16.6 CDD Human Microsome intrinsic clearance model built with data from AstraZeneca in ChEMBL showing the ROC plots for three fold cross validation.
Credit: European Bioinformatics Institute. DOI: 10.6084/m9.figshare.3206236.

in chemistry across the web.[15] Up until this point there was little interest in data quality.[98] Others have also recently compared bioactivity databases from commercial and public sources (ChEMBL, WOMBAT, PubChem, Evolvus and K_i Database) identifying errors such as incorrect molecular structures or stereoisomers in 8.8% of molecules.[11,12]

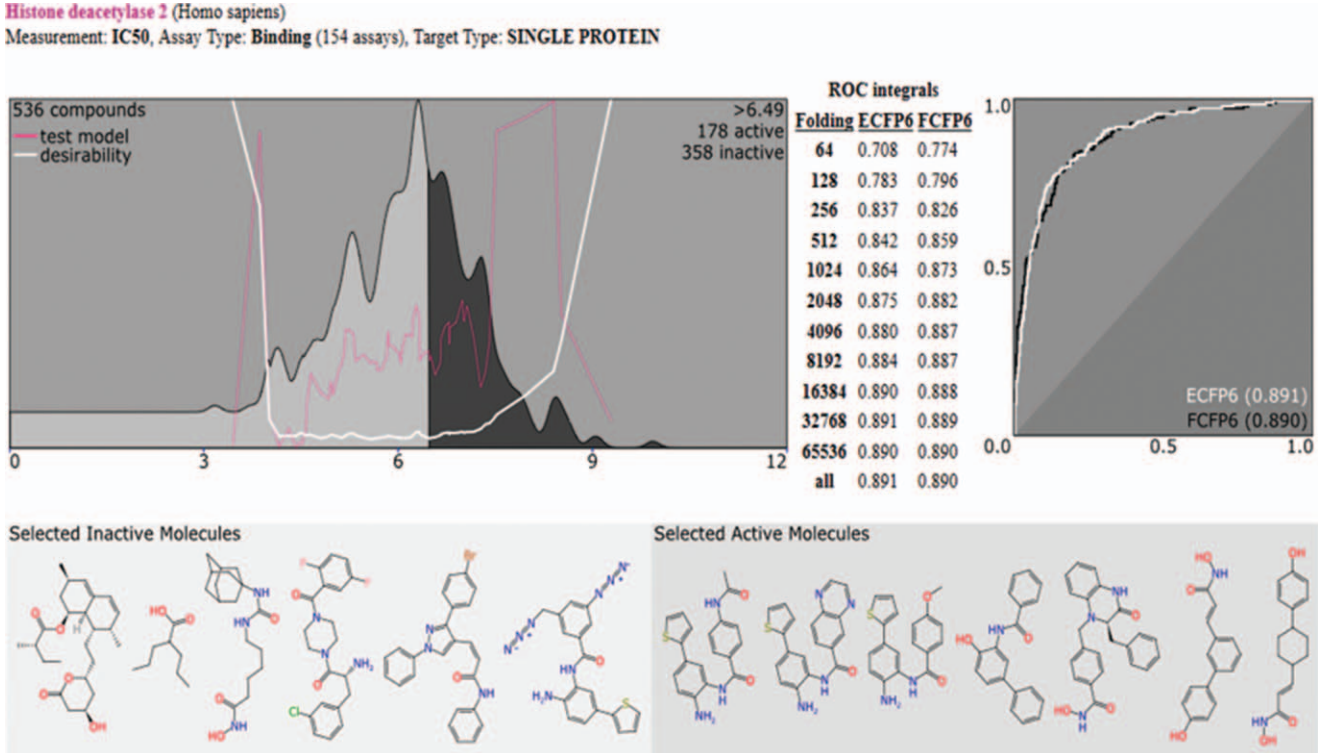


Figure 16.7 An example of an extracted dataset from ChEMBL and subsequent analysis, leading to the detection of a suitable activity threshold. This shows a plot of population *versus* activity, for which the solid curve shows the integral, which is colored to show inactive (below threshold: light grey) and active (above threshold: dark grey) molecules. The ROC integral for subset models at various thresholds is plotted, as is the overall desirability composite score. To the right is the ROC curve for ECFP6 and FCFP6 models, built using the whole dataset at the determined threshold. A representative diverse selection of ''active'' and ''inactive'' molecules is shown underneath.
Credit: European Bioinformatics Institute. DOI: 10.6084/m9.figshare.3206236.

Table 16.1 Bayesian models developed with ChEMBL and public data.

URL	Summary	Ref.
http://molsync.com/bayesian1	ADME/Tox and neglected disease datasets curated from public data	86
http://molsync.com/bayesian2	Models developed from ChEMBL	122
http://molsync.com/transporters	Select transporter datasets curated from public data	92
http://molsync.com/ebola/	Ebola models developed from curated data	84

Several researchers have also been drawn to the challenges in drug repurposing when molecule structures from industry with ambiguous identifiers are shared, resulting in analytical errors.[99] Similar problems arise when dealing with patents and massive data disclosures.[2] Some of these issues have also come to the fore with the recent clinical trial tragedy that resulted in serious adverse events and one death with BIA 10–2474, in which the structure and bioactivity data had not been disclosed. Initially, a particular structure was used with target prediction software, but was subsequently found to be the incorrect structure. The lack of mapping between patents and identifiers confounded the problem and it took a week of speculation before the structure's name was disclosed in a leaked protocol. Even then, 3 months elapsed before an official report with some new data surfaced, but we are still no closer to a causative understanding of the tragedy at the mechanistic toxicology level.[100–102]

Other technical issues that have surfaced over the years specifically relate to some of the aspects of generating bioactivity data in the first place. Even steps such as how a liquid is dispensed and how dilution steps are constructed have been found to have a profound effect on the bioactivity of a compound.[103] These types of potential sources of error can also to some extent be modeled mathematically,[104] suggesting perhaps that we could correct data in databases if we had a complete understanding of how they were generated, including details such as what hardware was used to run the experiment. This points to the importance of complete documentation and creation of bioassay ontologies.[47,105–107] There are likely to be many efforts and companies that could exploit this important aspect to improve our current bioactivity databases.

There are certainly many other areas that could be improved, including ensuring that data from papers are automatically deposited in databases as a way to limit potential errors. Bioactivity data should move in a lossless manner *via* electronic formats, preferably using open community standards, rather than having a third party curation step.[108] Also, the deposition of bioactivity data (molecular structures, experimental protocols and activity values) should be considered as important as data types, such as crystal structures, and deposition should be mandated prior to publication. There are numerous standards that have been created that could be readily

followed, *e.g.* minimum information about a bioactive entity (MIABE).[109] Collective encouragement by publishers and/or requirements from funding organizations such as the NIH to require direct data deposition could help this happen, as was done for the deposition of the majority of the data points in PubChem. However, despite the successes of the Protein DataBank, GenBank and the Crystallographic Structure Database community, agreement in terms of the deposition of experimental data and descriptors of associated metadata, for example ADME/Tox data, has not come to fruition despite encouragement[110] and available platforms for hosting models.[95,111] Platforms and options already exist that could support the mandated deposition of bioactivity data. Time will tell whether this situation will change.

16.4 Conclusions

Bioactivity databases, both large and small, are a valuable asset for researchers working in drug discovery and other areas of the biomedical industry. We, and others, have illustrated how the curation of such data creates a starting point for large scale machine learning and target inference methods. At the same time, most of these databases do not provide data in a format that can be readily used for modeling so there is an opportunity for improvement. There are also few databases that allow the user to select the data to build their own machine learning models with either public data, their own data or a combination of the two. Of course the challenge here is testing the models and evaluating their applicability[112–117] in such a way that the user does not need extensive cheminformatics expertise. This is a tall order, especially considering how long it has taken us to get to where we are today. In addition, there is still naivety regarding what databases are available online, and difficulty understanding the complexities of their data structures in order to make informed judgments on quality, content and fitness for purposes. This chapter has hopefully introduced a few more databases to the reader of which they may not have been aware and clarified their role in the ecosystem of bioactivity databases. There are likely many more of interest to the readers in the Nucleic Acids Research database summary list.[118] However, there are still far too many flat files of data ''out in the wild'' that should either be meshed into new databases or preferably one of the existing databases such as CDD Public. In this way, these ''lost data'' can be made mineable and useful for modeling. The benefits of having access to thousands of machine learning models created from such data means that a scientist could start a new project, use the model to suggest new compounds to test and, with a collaborator, readily validate the predictions. From our experience, this is already feasible and the proof of concept has been repeatedly demonstrated.[84,119,120]

How do we justify the continued costs of creating and maintaining these databases? The success that has resulted from the development of these databases, or the data residing in them, should be continually highlighted (and if necessary celebrated) as it may result in additional usage or even

encourage data contribution. Any drug repositioning[121] opportunities that could be attributed to one or more databases should highlight the value and business proposition. We are not aware of this type of retrospective analysis of databases to assess their successes and perhaps this is long overdue. It is clear that without public bioactivity databases the researcher would be entirely dependent on commercial databases, which for many would be out of reach. Therefore, there should be a balance between making data (which in the majority of cases has been generated with public funding) generally accessible and providing incentives for companies to develop new software and database products. Freely accessible bioactivity databases fill a gap that existed over a decade ago, but their long term viability remains unclear and how we will use them in the next 5–10 years will depend on a combination of issues: data quality, data licenses and software tools for analysis, mining, modeling, and data distribution. Progress *is* being made in all of these areas and we should be optimistic, but it is likely important that we start collating examples to justify how increasingly limited research funding can have the highest impact with bioactivity databases.

Acknowledgements

We acknowledge that the Bayesian model software within CDD was developed with support from Award Number 9R44TR000942-02 ''Biocomputation across distributed private datasets to enhance drug discovery'' from the NIH NCATS. The CDD tuberculosis has been developed thanks to funding from the Bill and Melinda Gates Foundation (grant 49852 ''Collaborative drug discovery for TB through a novel database of SAR data optimized to promote data archiving and sharing''). The work was partially supported by a grant from the European Community's Seventh Framework Program (grant 260872, MM4TB Consortium). S. Ekins and B. A. Bunin sincerely acknowledge many colleagues, collaborators and advocates who have contributed to the development of CDD over the years.

References

1. A. J. Williams, *Curr. Opin. Drug Discovery Dev.*, 2008, **11**, 393–404.
2. C. A. Lipinski, N. Litterman, C. Southan, A. J. Williams, A. M. Clark and S. Ekins, *J. Med. Chem.*, 2015, **58**, 2068–2076.
3. M. A. Johnson and G. M. Maggiora, *Concepts and Applications of Molecular Similarity*, Wiley-Interscience, Hoboken, NY, 1990.
4. P. Willett, J. M. Barnard and G. M. Downs, *J. Chem. Inf. Comput. Sci.*, 1998, **38**, 983–996.
5. G. Patlewicz, N. Jeliazkova, A. Gallegos Saliner and A. P. Worth, *SAR QSAR Environ. Res.*, 2008, **19**, 397–412.
6. Y. Low, A. Sedykh, D. Fourches, A. Golbraikh, M. Whelan, I. Rusyn and A. Tropsha, *Chem. Res. Toxicol.*, 2013, **26**, 1199–1208.

7. H. Zhu, J. Zhang, M. T. Kim, A. Boison, A. Sedykh and K. Moran, *Chem. Res. Toxicol.*, 2014, **27**, 1643–1651.
8. J. Zhang, J. H. Hsieh and H. Zhu, *PLoS One*, 2014, **9**, e99863.
9. A. B. Wagner, *J. Chem. Inf. Model.*, 2006, **46**, 767–774.
10. C. Southan, P. Varkonyi and S. Muresan, *J. Cheminf.*, 2009, **1**, 10.
11. P. Tiikkainen and L. Franke, *J. Chem. Inf. Model.*, 2012, **52**, 319–326.
12. P. Tiikkainen, L. Bellis, Y. Light and L. Franke, *J. Chem. Inf. Model.*, 2013, **53**, 2499–2505.
13. A. J. Williams and S. Ekins, *Drug Discovery Today*, 2011, **16**, 747–750.
14. A. J. Williams, S. Ekins, O. Spjuth and E. L. Willighagen, *Methods Mol. Biol.*, 2012, **929**, 221–241.
15. A. J. Williams, S. Ekins and V. Tkachenko, *Drug Discovery Today*, 2012, **17**, 685–701.
16. N. K. Litterman and S. Ekins, *Drug Discovery Today*, 2014.
17. S. Ekins, J. S. Freundlich, I. Choi, M. Sarker and C. Talcott, *Trends Microbiol.*, 2011, **19**, 65–74.
18. C. Southan, M. Sitzmann and S. Muresan, *Mol. Inf.*, 2013, **32**, 881–897.
19. J. Rosen, J. Gottfries, S. Muresan, A. Backlund and T. I. Oprea, *J. Med. Chem.*, 2009, **52**, 1953–1962.
20. B. L. Roth, W. K. Kroeze, S. J. Patel and E. Lopez, *The Neuroscientist*, 2000, **6**, 252–262.
21. S. L. Mathias, J. Hines-Kay, J. J. Yang, G. Zahoransky-Kohalmi, C. G. Bologa, O. Ursu and T. I. Oprea, *Database*, 2013, **2013**, bat044.
22. J. Kringelum, S. K. Kjaerulff, S. Brunak, O. Lund, T. I. Oprea and O. Taboureau, *Database*, 2016, **2016**.
23. J. Hert, M. J. Keiser, J. J. Irwin, T. I. Oprea and B. K. Shoichet, *J. Chem. Inf. Model.*, 2008, **48**, 755–765.
24. F. Ntie-Kang, D. Zofou, S. B. Babiaka, R. Meudom, M. Scharfe, L. L. Lifongo, J. A. Mbah, L. M. Mbaze, W. Sippl and S. M. Efange, *PLoS One*, 2013, **8**, e78085.
25. K. Sanderson, *Nature*, 2011, **469**, 18–20.
26. A. M. Richard, R. S. Judson, K. A. Houck, C. M. Grulke, P. Volarath, I. Thillaonadarajah, C. Yang, J. Rathman, M. T. Martin, J. F. Wambaugh, T. B. Knudsen, J. Kancherla, K. Mansouri, G. Patlewicz, A. Williams, S. B. Little, K. M. Crofton and R. S. Thomas, *Submitted*, 2016.
27. Anon, https://www.epa.gov/chemical-research/toxicity-forecaster-toxcasttm-data, 2015.
28. R. Huang, M. Xia, S. Sakamuru, J. Zhao, S. A. Shahane, M. Attene-Ramos, T. Zhao, C. P. Austin and A. Simeonov, *Nat. Commun.*, 2016, **7**, 10425.
29. Anon, https://www.ncbi.nlm.nih.gov/pcassay?term=%22tox21%22.
30. J. Meslamani, S. G. Smith, R. Sanchez and M. M. Zhou, *Bioinformatics*, 2014, **30**, 1481–1483.
31. D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam and M. Hassanali, *Nucleic Acids Res.*, 2008, **36**, D901–906.

32. W. K. Chan, H. Zhang, J. Yang, J. R. Brender, J. Hur, A. Ozgur and Y. Zhang, *Bioinformatics*, 2015, **31**, 3035–3042.
33. R. L. Strausberg and S. L. Schreiber, *Science*, 2003, **300**, 294–295.
34. N. Tolliday, P. A. Clemons, P. Ferraiolo, A. N. Koehler, T. A. Lewis, X. Li, S. L. Schreiber, D. S. Gerhard and S. Eliasof, *Cancer Res.*, 2006, **66**, 8935–8942.
35. K. P. Seiler, G. A. George, M. P. Happ, N. E. Bodycombe, H. A. Carrinski, S. Norton, S. Brudz, J. P. Sullivan, J. Muhlich, M. Serrano, P. Ferraiolo, N. J. Tolliday, S. L. Schreiber and P. A. Clemons, *Nucleic Acids Res.*, 2008, **36**, D351–359.
36. S. Louise-May, B. Bunin and S. Ekins, *Touch Briefings – Drug Discovery*, 2009, **6**, 17–21.
37. A. J. Williams, V. Tkachenko, C. Lipinski, A. Tropsha and S. Ekins, *Drug Discovery World*, 2009, **10**, 33–38. Winter.
38. X. Chen, Y. Lin, M. Liu and M. K. Gilson, *Bioinformatics*, 2002, **18**, 130–139.
39. M. K. Gilson, T. Liu, M. Baitaluk, G. Nicola, L. Hwang and J. Chong, *Nucleic Acids Res.*, 2016, **44**, D1045–1053.
40. B. Chen, K. J. McConnell, N. Wale, D. J. Wild and E. M. Gifford, *Bioinformatics*, 2011, **27**, 3044–3049.
41. M. J. Keiser, B. L. Roth, B. N. Armbruster, P. Ernsberger, J. J. Irwin and B. K. Shoichet, *Nat. Biotechnol.*, 2007, **25**, 197–206.
42. M. J. Keiser, V. Setola, J. J. Irwin, C. Laggner, A. I. Abbas, S. J. Hufeisen, N. H. Jensen, M. B. Kuijer, R. C. Matos, T. B. Tran, R. Whaley, R. A. Glennon, J. Hert, K. L. Thomas, D. D. Edwards, B. K. Shoichet and B. L. Roth, *Nature*, 2009, **462**, 175–181.
43. S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker, J. Wang, B. Yu, J. Zhang and S. H. Bryant, *Nucleic Acids Res.*, 2016, **44**, D1202–1213.
44. Y. Wang, J. Xiao, T. O. Suzek, J. Zhang, J. Wang and S. H. Bryant, *Nucleic Acids Res.*, 2009, **37**, W623–633.
45. Y. Wang, E. Bolton, S. Dracheva, K. Karapetyan, B. A. Shoemaker, T. O. Suzek, J. Wang, J. Xiao, J. Zhang and S. H. Bryant, *Nucleic Acids Res.*, 2010, **38**, D255–266.
46. N. Litterman, C. A. Lipinski, B. A. Bunin and S. Ekins, *J. Chem. Inf. Model.*, 2014, **54**, 2996–3004.
47. E. A. Howe, A. de Souza, D. L. Lahr, S. Chatwin, P. Montgomery, B. R. Alexander, D. T. Nguyen, Y. Cruz, D. A. Stonich, G. Walzer, J. T. Rose, S. C. Picard, Z. Liu, J. N. Rose, X. Xiang, J. Asiedu, D. Durkin, J. Levine, J. J. Yang, S. C. Schurer, J. C. Braisted, N. Southall, M. R. Southern, T. D. Chung, S. Brudz, C. Tanega, S. L. Schreiber, J. A. Bittker, R. Guha and P. A. Clemons, *Nucleic Acids Res.*, 2015, **43**, D1163–1170.
48. Anon, https://pubchem.ncbi.nlm.nih.gov/lcss/, 2011.
49. Anon, https://pubchem.ncbi.nlm.nih.gov/classification/#hid=72.
50. Anon, https://pubchem.ncbi.nlm.nih.gov/widget/docs/widget_help.html#bioactivity, 2012.

51. H. E. Pence and A. J. Williams, *J. Chem. Educ.*, 2010, **87**, 1123–1124.
52. T. Cheng, Q. Li, Y. Wang and S. H. Bryant, *J. Chem. Inf. Model.*, 2011, **51**, 2440–2448.
53. A. J. Williams, J. Wilbanks and S. Ekins, *PLoS Comput. Biol.*, 2012, **8**, e1002706.
54. M. P. Gleeson, A. Hersey, D. Montanari and J. Overington, *Nat. Rev. Drug Discovery.*, 2011, **10**, 197–208.
55. A. P. Bento, A. Gaulton, A. Hersey, L. J. Bellis, J. Chambers, M. Davies, F. A. Kruger, Y. Light, L. Mak, S. McGlinchey, M. Nowotka, G. Papadatos, R. Santos and J. P. Overington, *Nucleic Acids Res.*, 2014, **42**, D1083–1090.
56. G. Papadatos and J. P. Overington, *Future Med. Chem.*, 2014, **6**, 361–364.
57. J. Chambers, http://chembl.blogspot.com/2011/09/integration-of-filtered-set-of-pubchem.html, 2011.
58. J. Tang, A. Szwajda, S. Shakyawar, T. Xu, P. Hintsanen, K. Wennerberg and T. Aittokallio, *J. Chem. Inf. Model.*, 2014, **54**, 735–743.
59. G. J. van Westen, O. O. van den Hoven, R. van der Pijl, T. Mulder-Krieger, H. de Vries, J. K. Wegner, A. P. Ijzerman, H. W. van Vlijmen and A. Bender, *J. Med. Chem.*, 2012, **55**, 7010–7020.
60. F.-J. Gamo, L. M. Sanz, J. Vidal, C. de Cozar, E. Alvarez, J.-L. Lavandera, D. E. Vanderwall, D. V. S. Green, V. Kumar, S. Hasan, J. R. Brown, C. E. Peishoff, L. R. Cardon and J. F. Garcia-Bustos, *Nature*, 2010, **465**, 305–310.
61. M. C. Wenlock, R. P. Austin, P. Barton, A. M. Davis and P. D. Leeson, *J. Med. Chem.*, 2003, **46**, 1250–1256.
62. L. J. Bellis, R. Akhtar, B. Al-Lazikani, F. Atkinson, A. P. Bento, J. Chambers, M. Davies, A. Gaulton, A. Hersey, K. Ikeda, F. A. Kruger, Y. Light, S. McGlinchey, R. Santos, B. Stauch and J. P. Overington, *Biochem. Soc. Trans.*, 2011, **39**, 1365–1370.
63. C. Southan, J. L. Sharman, H. E. Benson, E. Faccenda, A. J. Pawson, S. P. Alexander, O. P. Buneman, A. P. Davenport, J. C. McGrath, J. A. Peters, M. Spedding, W. A. Catterall, D. Fabbro, J. A. Davies and I. Nc, *Nucleic Acids Res.*, 2016, **44**, D1054–1068.
64. A. J. Pawson, J. L. Sharman, H. E. Benson, E. Faccenda, S. P. Alexander, O. P. Buneman, A. P. Davenport, J. C. McGrath, J. A. Peters, C. Southan, M. Spedding, W. Yu, A. J. Harmar and I. Nc, *Nucleic Acids Res.*, 2014, **42**, D1098–1106.
65. S. Ekins, M. Hohman and B. A. Bunin, in *Collaborative Computational Technologies for Biomedical Research*, ed. S. Ekins, M. A. Z. Hupcey and A. J. Williams, Wiley and Sons, Hoboken, 2011, vol. 335–361.
66. M. Hohman, K. Gregory, K. Chibale, P. J. Smith, S. Ekins and B. Bunin, *Drug Discovery Today*, 2009, **14**, 261–270.
67. S. Ekins and B. A. Bunin, *Methods Mol. Biol.*, 2013, **993**, 139–154.
68. S. Ekins, J. Bradford, K. Dole, A. Spektor, K. Gregory, D. Blondeau, M. Hohman and B. Bunin, *Mol. BioSyst.*, 2010, **6**, 840–851.

69. S. Ekins, T. Kaneko, C. A. Lipinksi, J. Bradford, K. Dole, A. Spektor, K. Gregory, D. Blondeau, S. Ernst, J. Yang, N. Goncharoff, M. Hohman and B. Bunin, *Mol. BioSyst.*, 2010, **6**, 2316–2324.
70. S. Ekins and J. S. Freundlich, *Pharm. Res.*, 2011, **28**, 1859–1869.
71. G. Lamichhane, J. S. Freundlich, S. Ekins, N. Wickramaratne, S. Nolan and W. R. Bishai, *MBio*, 2011, **2**, e00301–00310.
72. M. Sarker, C. Talcott, P. Madrid, S. Chopra, B. A. Bunin, G. Lamichhane, J. S. Freundlich and S. Ekins, *Pharm. Res.*, 2012, **29**, 2115–2127.
73. S. Ekins, J. S. Freundlich and R. C. Reynolds, *J. Chem. Inf. Model.*, 2013, **53**, 3054–3063.
74. S. Ekins, J. S. Freundlich and R. C. Reynolds, *Figshare*, 2013.
75. S. Ekins, R. C. Reynolds, S. G. Franzblau, B. Wan, J. S. Freundlich and B. A. Bunin, *PLoS One*, 2013, **8**, e63240.
76. S. Ekins, A. C. Casey, D. Roberts, T. Parish and B. A. Bunin, *Tuberculosis (Edinb)*, 2014, **94**, 162–169.
77. S. Ekins, J. S. Freundlich, J. V. Hobrath, E. Lucile White and R. C. Reynolds, *Pharm. Res.*, 2014, **31**, 414–435.
78. S. Ekins, J. S. Freundlich and R. C. Reynolds, *J. Chem. Inf. Model.*, 2014, **54**, 2157–2165.
79. S. Ekins, E. L. Nuermberger and J. S. Freundlich, *Drug Discovery Today*, 2014, **19**, 1279–1282.
80. S. Ekins, R. Pottorf, R. C. Reynolds, A. J. Williams, A. M. Clark and J. S. Freundlich, *J. Chem. Inf. Model.*, 2014, **54**, 1070–1082.
81. S. E. Ekins, P. E. Madrid, M. Sarker, S.-G. Li, N. Mittal, P. Kumar, X. Wang, T. P. Stratton, M. Zimmerman, C. Talcott, P. Bourbon, M. Travers, M. Yadav and J. S Freundlich, *PLoS One*, 2015, **10**, e0141076.
82. L. Zhang, D. Fourches, A. Sedykh, H. Zhu, A. Golbraikh, S. Ekins, J. Clark, M. C. Connelly, M. Sigal, D. Hodges, A. Guiguemde, R. K. Guy and A. Tropsha, *J. Chem. Inf. Model.*, 2013, **53**, 475–492.
83. S. Ekins, J. L. de Siqueira-Neto, L. I. McCall, M. Sarker, M. Yadav, E. L. Ponder, E. A. Kallel, D. Kellar, S. Chen, M. Arkin, B. A. Bunin, J. H. McKerrow and C. Talcott, *PloS Neglected Trop. Dis.*, 2015, **9**, e0003878.
84. S. Ekins, J. S. Freundlich, A. M. Clark, M. Anantpadma, R. A. Davey and P. Madrid, *F1000Res*, 2016, **4**, 1091.
85. S. Ekins, A. M. Clark and M. Sarker, *J. Cheminf.*, 2013, **5**, 13.
86. A. M. Clark, K. Dole, A. Coulon-Spector, A. McNutt, G. Grass, J. S. Freundlich, R. C. Reynolds and S. Ekins, *J. Chem. Inf. Model.*, 2015, **55**, 1231–1245.
87. R. R. Gupta, E. M. Gifford, T. Liston, C. L. Waller, B. Bunin and S. Ekins, *Drug Metab. Dispos.*, 2010, **38**, 2083–2090.
88. S. Ekins, A. C. Casey, D. Roberts, T. Parish and B. A. Bunin, *Tuberculosis (Edinb)*, 2014, **94**, 162–169.
89. S. Ekins, R. C. Reynolds, H. Kim, M. S. Koo, M. Ekonomidis, M. Talaue, S. D. Paget, L. K. Woolhiser, A. J. Lenaerts, B. A. Bunin, N. Connell and J. S. Freundlich, *Chem. Biol.*, 2013, **20**, 370–378.

90. S. Ekins, J. S. Freundlich, J. V. Hobrath, E. L. White and R. C. Reynolds, *Pharm. Res.*, 2014, **31**, 414–435.
91. A. L. Perryman, T. P. Stratton, S. Ekins and J. S. Freundlich, *Pharm. Res.*, 2016, **33**, 433–449.
92. S. Ekins, A. M. Clark and S. H. Wright, *Drug Metab. Dispos.*, 2015, **43**, 1642–1645.
93. A. M. Clark, K. Dole and S. Ekins, *J. Chem. Inf. Model.*, 2015, **56**, 275–285.
94. T. Walker, C. M. Grulke, D. Pozefsky and A. Tropsha, *Bioinformatics*, 2010, **26**, 3000–3001.
95. I. Sushko, S. Novotarskyi, R. Korner, A. K. Pandey, M. Rupp, W. Teetz, S. Brandmaier, A. Abdelaziz, V. V. Prokopenko, V. Y. Tanchuk, R. Todeschini, A. Varnek, G. Marcou, P. Ertl, V. Potemkin, M. Grishina, J. Gasteiger, C. Schwab, V. A. Baskin II, E. V. Palyulin, W. J. Radchenko, V. Welsh, D. Kholodovych, A. Chekmarev, J. Cherkasov, Q. Y. Aires-de-Sousa, A. Zhang, F. Bender, L. Nigsch, A. Patiny, V. Williams, Tkachenko and I. V. Tetko, *J. Comput.-Aided Mol. Des.*, 2011, **25**, 533–554.
96. A. M. Clark, M. Sarker and S. Ekins, *J. Cheminf.*, 2014, **6**, 38.
97. S. Ekins and A. J. Williams, *Drug Discovery Today.*, 2010, **15**, 812–815.
98. D. Fourches, E. Muratov and A. Tropsha, *J. Chem. Inf. Model.*, 2010, **50**, 1189–1204.
99. C. Southan, A. J. Williams and S. Ekins, *Drug Discovery Today.*, 2013, **18**, 58–70.
100. C. Southan, http://cdsouthan.blogspot.com/2016/01/the-unfortunate-case-of-bia-10-2474.html, 2016.
101. A. J. Williams, http://www.chemconnector.com/2016/01/24/bia-10-2474-confusions-in-chemical-structure-and-the-need-for-early-clarity-in-chemical-structures/, 2016.
102. S. Ekins, http://www.collabchem.com/2016/01/16/what-can-we-predict-about-bia-10-2474/, 2016.
103. S. Ekins, J. Olechno and A. J. Williams, *PLoS One*, 2013, **8**, e62325.
104. S. M. Hanson, S. Ekins and J. D. Chodera, *J. Comput.-Aided Mol. Des.*, 2015, **29**, 1073–1086.
105. A. M. Clark, B. A. Bunin, N. K. Litterman, S. C. Schurer and U. Visser, *PeerJ*, 2014, **2**, e524.
106. U. Visser, S. Abeyruwan, U. Vempati, R. P. Smith, V. Lemmon and S. C. Schurer, *BMC Bioinf.*, 2011, **12**, 257.
107. A. de Souza, J. A. Bittker, D. L. Lahr, S. Brudz, S. Chatwin, T. I. Oprea, A. Waller, J. J. Yang, N. Southall, R. Guha, S. C. Schurer, U. D. Vempati, M. R. Southern, E. S. Dawson, P. A. Clemons and T. D. Chung, *J. Biomol. Screening*, 2014, **19**, 614–627.
108. A. M. Clark, A. J. Williams and S. Ekins, *J. Cheminf.*, 2015, **7**, 9.
109. S. Orchard, B. Al-Lazikani, S. Bryant, D. Clark, E. Calder, I. Dix, O. Engkvist, M. Forster, A. Gaulton, M. Gilson, R. Glen, M. Grigorov, K. Hammond-Kosack, L. Harland, A. Hopkins, C. Larminie, N. Lynch, R. K. Mann, P. Murray-Rust, E. Lo Piparo, C. Southan, C. Steinbeck,

D. Wishart, H. Hermjakob, J. Overington and J. Thornton, *Nat. Rev. Drug Discovery*, 2011, **10**, 661–669.
110. S. Ekins and A. J. Williams, *Lab Chip*, 2010, **10**, 13–22.
111. V. Ruusmann, S. Sild and U. Maran, *J. Cheminf.*, 2015, **7**, 32.
112. I. V. Tetko, I. Sushko, A. K. Pandey, H. Zhu, A. Tropsha, E. Papa, T. Oberg, R. Todeschini, D. Fourches and A. Varnek, *J. Chem. Inf. Model.*, 2008, **48**, 1733–1746.
113. A. Tropsha and A. Golbraikh, *Curr. Pharm. Des.*, 2007, **13**, 3494–3504.
114. D. W. Roberts, G. Patlewicz, P. S. Kern, F. Gerberick, I. Kimber, R. J. Dearman, C. A. Ryan, D. A. Basketter and A. O. Aptula, *Chem. Res. Toxicol.*, 2007, **20**, 1019–1030.
115. I. V. Tetko, P. Bruneau, H. W. Mewes, D. C. Rohrer and G. I. Poda, *Drug Discovery Today*, 2006, **11**, 700–707.
116. S. Dimitrov, G. Dimitrova, T. Pavlov, N. Dimitrova, G. Patlewicz, J. Niemela and O. Mekenyan, *J. Chem. Inf. Model.*, 2005, **45**, 839–849.
117. R. P. Sheridan, *J. Chem. Inf. Model.*, 2013, **53**, 2837–2850.
118. Anon, https://www.oxfordjournals.org/our_journals/nar/database/a/, 2014.
119. S. Ekins, R. Reynolds, H. Kim, M.-S. Koo, M. Ekonomidis, M. Talaue, S. D. Paget, L. K. Woolhiser, A. J. Lenaerts, B. A. Bunin, N. Connell and J. S. Freundlich, *Chem. Biol.*, 2013, **20**, 370–378.
120. S. Ekins, J. Lage de Siqueira-Neto, L.-I. McCall, M. Sarker, M. Yadav, E. L. Ponder, E. A. Kallel, D. Kellar, S. Chen, M. Arkin, B. A. Bunin, J. H. McKerrow and C. Talcott, *PloS Neglected Trop. Dis.*, 2015, **9**, e0003878.
121. S. Ekins, A. J. Williams, M. D. Krasowski and J. S. Freundlich, *Drug Discovery Today*, 2011, **16**, 298–310.
122. A. M. Clark and S. Ekins, *J. Chem. Inf. Model.*, 2015, **55**, 1246–1260.